Jailbreaking Attack against Multimodal Large Language Model

Year: 2024 | *Citations: 112* | *Authors: Zhenxing Niu, Haodong Ren, Xinbo Gao*

Abstract

This paper focuses on jailbreaking attacks against multi-modal large language models (MLLMs), seeking to elicit MLLMs to generate objectionable responses to harmful user queries. A maximum likelihood-based algorithm is proposed to find an \emph{image Jailbreaking Prompt} (imgJP), enabling jailbreaks against MLLMs across multiple unseen prompts and images (i.e., data-universal property). Our approach exhibits strong model-transferability, as the generated imgJP can be transferred to jailbreak various models, including MiniGPT-v2, LLaVA, InstructBLIP, and mPLUG-Owl2, in a black-box manner. Moreover, we reveal a connection between MLLM-jailbreaks and LLM-jailbreaks. As a result, we introduce a construction-based method to harness our approach for LLM-jailbreaks, demonstrating greater efficiency than current state-of-the-art methods. The code is available here. \textbf{Warning: some content generated by language models may be offensive to some readers.}